COMPARING PREDICTIONS OF LEXICAL NORM DATA OBTAINED USING WORD ASSOCIATIONS AND WORD CO-OCCURENCE

Hendrik Vankrunkelsven, Steven Verheyen, Simon De Deyne, Gert Storms
Brain and Cognition, University of Leuven, Belgium

1. Introduction

Predicting lexical norm data has been done via **text corpora** (e.g., Bestgen & Vincze, 2012; Recchia & Louwerse, 2014) and a **word association corpus** (Vankrunkelsven, Verheyen, De Deyne, & Storms, 2015)

We **compare** the quality of prediction using both sources of data.

We predict 3 affective variables: **valence**, **dominance**, and **arousal**

And 2 non-affective variables: **concreteness**, and **age of acquisition** (**AoA**)

These predictions are cross-validated using lexical norm data

2. Method

2.1. Data

Text corpus:

Syntactic dependency model (De Deyne, Verheyen, & Storms, 2015) :

- Dutch articles in newspapers and magazines
- Internet web pages
- Dutch movie subtitles and Corpus of Spoken Dutch

103,842 lemma types

Similarities (cosine measure)

Word association corpus:

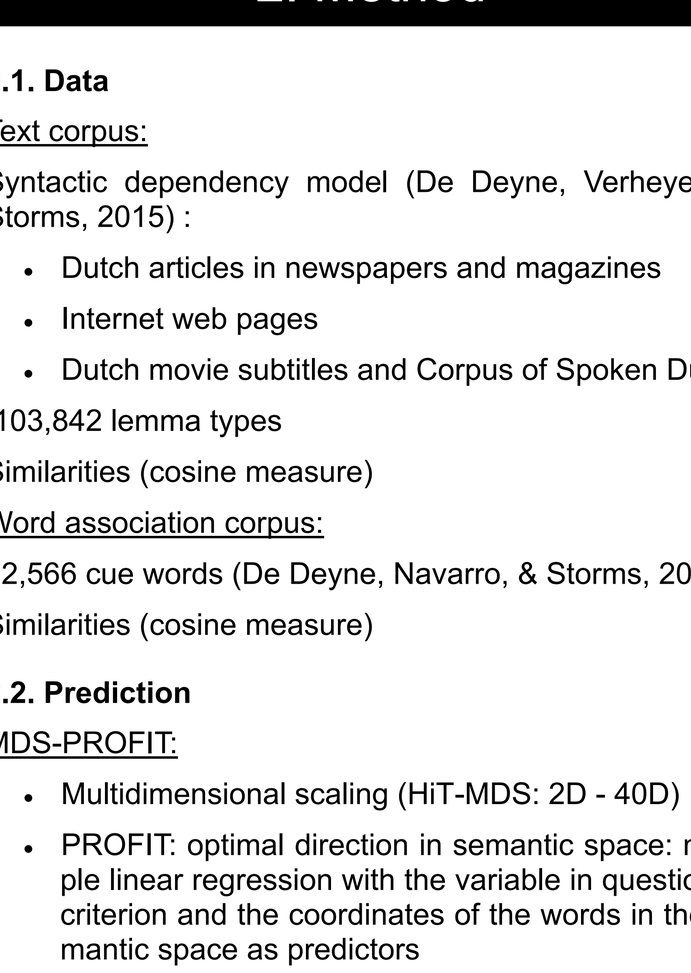
12,566 cue words (De Deyne, Navarro, & Storms, 2013)

Similarities (cosine measure)

2.2. Prediction

MDS-PROFIT:

- Multidimensional scaling (HiT-MDS: 2D - 40D)
- PROFIT: optimal direction in semantic space: multiple linear regression with the variable in question as criterion and the coordinates of the words in the semantic space as predictors
- Prediction word(s): projection(s) on this optimal direction



Example of a 2 dimensional semantic space consisting of 20 words. Words with green dots (10) are used to determine the optimal direction, the remaining 10 words with red dots are projected on this line

K-nearest neighbors

Average of norm scores of variable that is predicted from the K-nearest neighbors (based on similarity). K: 1-50, 60, 70, 80, 90, 100

K-nearest neighbors weighted

Weighted average according to similarity

2.3. Validation

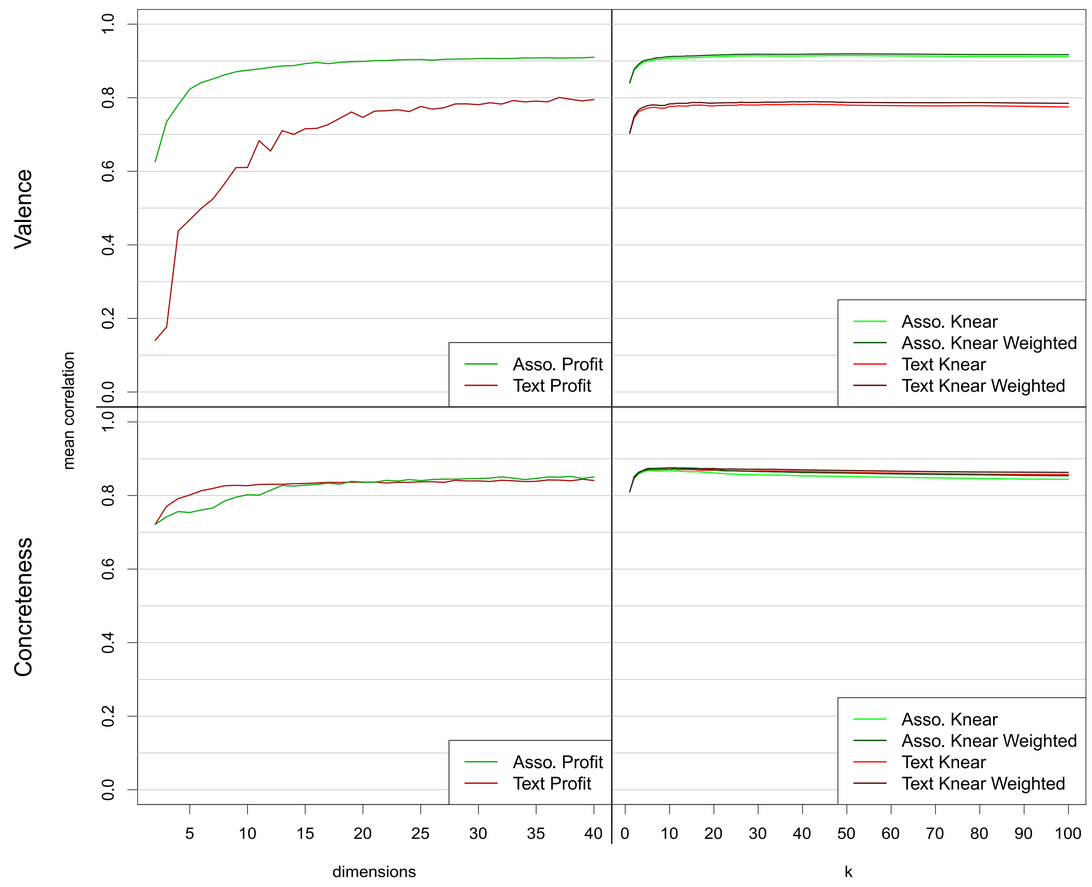
Cross-validation leave-one-out (L1O)

Prediction using every word except predicted word.

Vary size training/test set

To probe effect of size training set

3. Results



Cross-validation L1O Predictions

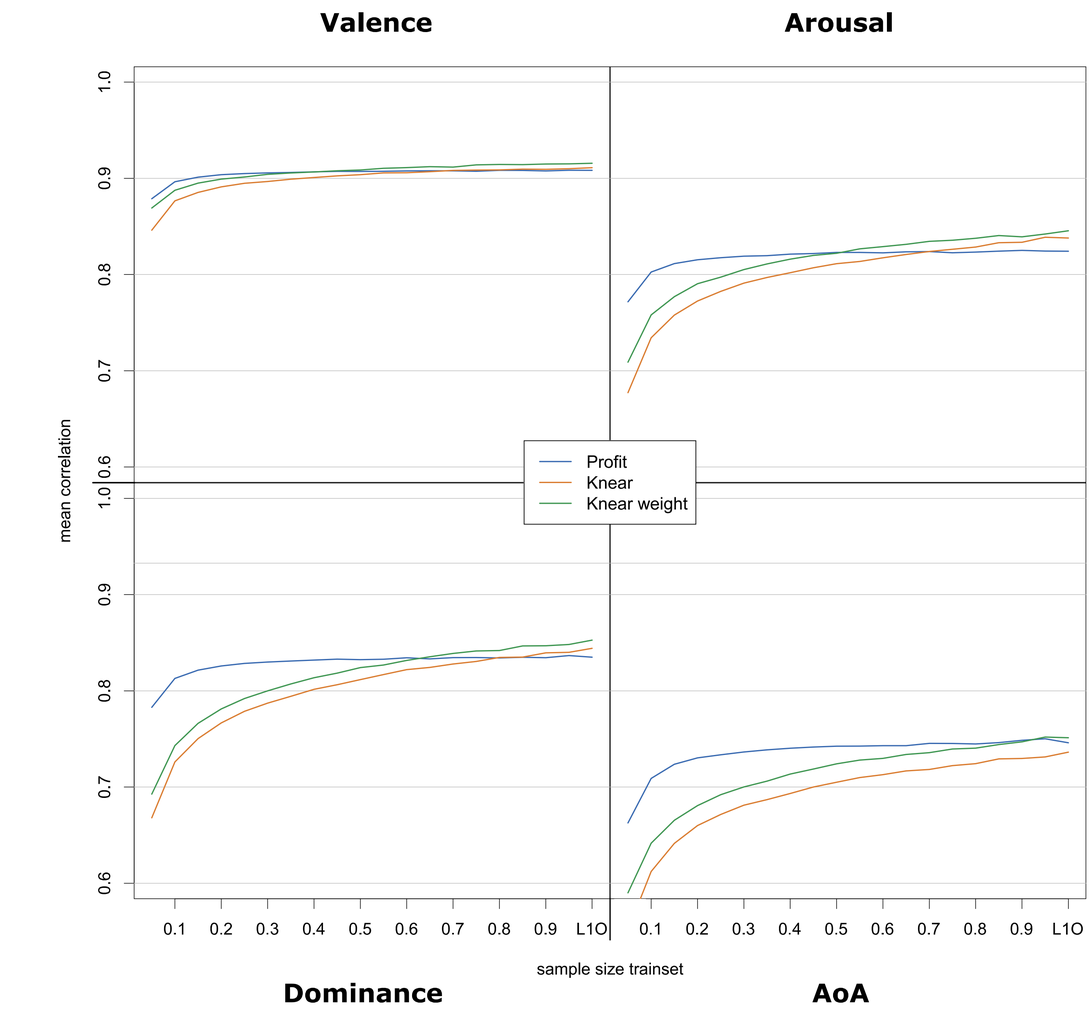
Cross-validation with norms for **valence**, arousal, dominance, and AoA from Moors et. al. (2013) and with norms for **concreteness** from Brysbaert et. al. (2014)

2831 words (present in all datasets)

Affective variables (valence, arousal, dominance) are **better** predicted from **association data**

Concreteness has same quality of prediction using text or associations

Weighted K-nearest neighbors gives best predictions with leave-one-out validation



Vary size training/test set

Only **association** data

3788 words (present in both datasets)

Training set between **5%** and **95%** of complete dataset (steps: of 5%)

For each variable and method optimal number of dimensions (2 to 50) and K from leave-one-out validation. (PROFIT: 49, 45, 49, 50; K-nearest: 50, 10, 10, 50; K-nearest weighted: 50, 13, 10, 50; for valence, arousal, dominance, AoA respectively)

PROFIT extrapolations **better** used with **smaller training sets** for arousal, dominance, AoA (and concreteness)

Weighted K-nearest neighbors consistently better then mere average of K-nearest neighbors

3. Results

Var:		Val.	Aro.	Dom.	AoA	Con.
PROFIT	Asso.	.91 [40]	.82 [38]	.83 [40]	.72 [39]	.85 [38]
	Text	.80 [37]	.70 [40]	.62 [40]	.73 [38]	.84 [39]
K-near	Asso.	.91 [50]	.84 [19]	.84 [8]	.71 [43]	.87 [10]
	Text	.78 [38]	.73 [8]	.66 [8]	.64 [24]	.87 [11]
K-near W.	Asso.	.92 [50]	.85 [19]	.85 [8]	.73 [48]	.88 [10]
	Text	.79 [43]	.74 [8]	.67 [8]	.64 [24]	.88 [11]

Best prediction [used dimension or K] for valence (Val.), arousal (Aro.), dominance (Dom.), age of acquisition (AoA), and concreteness (Con.) using association data (Asso.) or tekst data (Text)

References

Bestgen, Y., & Vincze, N. (2012). Checking and bootstrapping lexical norms by means of word similarity indexes. *Behavior Research Methods, 44*, 998–1006.

Brysbaert, M., Stevens, M., De Deyne, S., Voorspoels, W., & Storms, G. (2014). Norms of age of acquisition and concreteness for 30,000 Dutch words. *Acta Psychologica, 150*, 80–84.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods, 45*, 480–498.

De Deyne, S., Verheyen, S., & Storms, G. (2015). The role of corpus size and syntax in deriving lexico-semantic representations for a wide range of concepts. *Quarterly Journal of Experimental Psychology*, *68*, 1643–1664.

Moors, A., De Houwer, J., Hermans, D., Wanmaker, S., van Schie, K., Van Harmelen, A.-L., … Brysbaert, M. (2013). Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. *Behavior Research Methods, 45*, 169–177.

Recchia, G., & Louwerse, M. M. (2014). Reproducing affective norms with lexical co-occurrence statistics: Predicting valence, arousal, and dominance. *The Quarterly Journal of Experimental Psychology, 68*, 1–15.

Vankrunkelsven, H., Verheyen, S., De Deyne, S., Storms, G. (2015). Predicting lexical norms using a word association corpus. In Noelle, D. (Ed.), Dale, R. (Ed.), Warlaumont, A. (Ed.), Yoshimi, J. (Ed.), Matlock, T. (Ed.), Jennings, C. (Ed.), Maglio, P. (Ed.), Proceedings of the 37th Annual Conference of the Cognitive Science Society. Annual Conference of the Cognitive Science Society. Pasadena, CA, USA, 23-25 July 2015 (pp. 2463-2468).

Further Information

Contact **hendrik.vankrunkelsven@kuleuven.be**

KU LEUVEN

Sponsored by Academische Stichting Leuven